Vol.2016-DBS-163 No.5
Vol.2016-IFAT-123 No.5
2016/9/13

確率的データストリームにおける情報利得を用いたパターン照合手法

杉浦 健人[1,a)]　石川 佳治[1,b)]

概要：近年では大量のセンシングデータを収集し，機械学習技術を用いてそれらを解析する動きが盛んである．センシングデータを機械学習により解析したとき，結果は不確実性を含む**確率的データストリーム**となる．そこで，本稿では確率的データストリームに対して**パターン照合**を適用し，適切な照合結果（マッチ）の取得を目指す．確率的データストリームでは，クリーネ閉包を含むパターンを照合したとき，単純に確率の大きいマッチを検出するだけでは適切なマッチが得られない点が問題となる．本稿ではマッチの適切さを表す指標としてマッチの**情報利得**を定義し，情報利得を最大化することで適切なマッチを検出する手法を提案する．また，no-overlap セマンティクスを用いることで，Viterbi アルゴリズムにより効率的なパターン照合が行えることを示す．最後に，実データ及び人工データに基づく実験により，提案手法の有効性と効率性を示す．

1. はじめに

大量のセンシングデータが収集可能になったことで，機械学習技術によりそれらを活用しようとする動きが盛んである．近年ではスマートフォンなどだけでなく，Internet of Things（IoT）のようなモノからのセンシングも注目を浴びている．また，TensorFlow を始めとする機械学習ライブラリの公開により，集めたセンシングデータの解析も容易になった．センシングデータの解析による人の行動認識 [2,12] などはこれまで主に研究として行われてきたが，今後より実用的な場面で活用されることが予想される．

しかし，センシングデータを機械学習により解析したとき，結果が不確実性を含む**確率的データストリーム**となる点に注意しなければならない．機械学習による分類の結果としては，算出された確信度が最大のものを選択するのが一般的である．しかし，センシングデータにノイズがある場合やそもそも分類の境界が曖昧な場合など，複数の分類結果が同じような確信度を持つことは多々ある．このとき，単純に唯一の分類結果を選択するよりも，全ての分類結果の確信度を確率として表すほうが適切である．例えば，RFID による人の屋内位置のモニタリングを考える．RFID リーダが廊下にのみ設置されているとき，人が部屋の中にいることはわかるが，いずれの部屋にいるかはわからないことがある．このような場合，確率が最大の位置のみを選択するよりも，**図 1** のように各位置にいる確率を表した方がセンシングの結果を正確に反映できる．

event symbol	time step								
	1	2	3	4	5	6	7	8	9
RoomA (a)	0.5	0.4	0.5	0.1	0.1	0.1	0.4	0.5	0.4
Hall (h)	0.1	0.1	0.1	0.8	0.8	0.8	0.1	0.1	0.1
RoomB (b)	0.4	0.5	0.4	0.1	0.1	0.1	0.5	0.4	0.5

図 1　確率的データストリーム

本稿では確率的データストリームに対して**パターン照合**を適用し，適切な照合結果（マッチ）を得ることを目的とする．確率的データストリームに対する既存研究では，選択や射影，集約 [3,11]，Top-k 問合せ [6]，頻出アイテムの検出 [13]，クラスタリング [1] などが行われている．しかし，これらの研究では「部屋 A に留まっている」などの連続的なイベントの生起を効率的に検出できない．そこで本稿では，正規表現に基づくパターンを用いて，そうした連続的なイベント生起の効率的な検出を目指す．

確率的データストリームにパターン照合を適用したとき，マッチの適切さをどのように値で表すかが問題となる．既存研究 [7] では適切さの尺度としてマッチの生起確率を用いているが，これはクリーネ閉包を含むパターンには適用できない．例として図 1 へパターン $p = \langle \mathtt{a}^+ \rangle$ の照合を行った場合を考える．マッチの生起確率はマッチが示すイベントの同時確率であるため，生起確率が最も大きくなるのは $\langle \mathtt{a}_1 \rangle, \langle \mathtt{a}_3 \rangle, \langle \mathtt{a}_8 \rangle$ の三つのマッチである．しかし，

[1] 名古屋大学大学院情報科学研究科
[a)] sugiura@db.ss.is.nagoya-u.ac.jp
[b)] ishikawa@is.nagoya-u.ac.jp

$p = \langle \mathtt{a}^+ \rangle$ が与えられたとき，ユーザが実際に望む結果は $\langle \mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3 \rangle, \langle \mathtt{a}_6, \mathtt{a}_7, \mathtt{a}_8 \rangle$ のようなマッチだと考えられる．つまり，生起確率ではクリーネ閉包が表す「連続したマッチの生起」という意図を適切に評価できない．

そこで，マッチの適切さを表すための指標として**情報利得**（information gain）を提案する．まず，あるマッチ m に先験的な確率を与える関数 $\theta(m)$ を，予備実験などに基づきユーザが設計したとする．このとき，マッチ m が適切であるかは，

$$P(m) \geq \theta(m) \tag{1}$$

という式で判定できる．この式を変形すると，

$$\begin{aligned} \frac{P(m)}{\theta(m)} &\geq 1 \\ \log_2 \frac{P(m)}{\theta(m)} &\geq 0 \end{aligned} \tag{2}$$

となり，左辺に情報理論における情報利得の式が現れる．そこで，本稿では式 (2) 左辺をパターン照合における情報利得 $IG(m)$ として定義し，マッチの適切さの尺度に用いる．例として，ユーザがマッチの生起確率の推定に $\theta(m) = 0.2^{|m|}$ を用いる場合を考える．なお，$|m|$ は $|\langle \mathtt{a}_1, \mathtt{a}_2 \rangle| = 2$ のようにマッチ m の長さを表す．このとき，$m_1 = \langle \mathtt{a}_1 \rangle, m_2 = \langle \mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3 \rangle, m_3 = \langle \mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3, \mathtt{a}_4 \rangle$ の情報利得はそれぞれ以下のようになる．

$$\begin{aligned} IG(m_1) &= \log_2 \frac{0.5}{0.2} \simeq 1.32 \\ IG(m_2) &= \log_2 \frac{0.5 \cdot 0.4 \cdot 0.5}{0.2 \cdot 0.2 \cdot 0.2} \simeq 3.64 \\ IG(m_3) &= \log_2 \frac{0.5 \cdot 0.4 \cdot 0.5 \cdot 0.1}{0.2 \cdot 0.2 \cdot 0.2 \cdot 0.2} \simeq 2.64 \end{aligned}$$

長さが 1 しかない m_1 や生起確率の小さい “$\mathtt{a}_4$” を含む m_3 よりも m_2 の方が情報利得が大きく，連続して生起するイベントの適切さを上手く評価できていることがわかる．

本稿の構成は以下のとおりである．まず，2 章で関連研究を紹介し，3 章で入力となるパターンなどの定義を行う．その後，4 章で本稿で扱う問題を定義し，5 章でパターン照合のアルゴリズムについて具体的に述べる．最後に，6 章で提案手法を実験により評価し，7 章でまとめを述べる．

2. 関連研究

確率的データストリームに対してパターン照合を行う研究として，イベント生起のマルコフ性を考慮したものがある [8]．部屋 A から部屋 B へ移動するのに廊下を通らねばならないとき，1 タイムステップで部屋 A から部屋 B へ移動することは難しい．文献 [8] では，このようなイベント生起の相関をマルコフ性，つまりイベントの事後確率を用いることで表している．つまり，先ほどの例であれば，事後確率 $P(\mathtt{b}_t | \mathtt{a}_{t-1}) \simeq 0$ を用いることでそのようなイベント生起がほとんどないことを表す．しかしこの研究では，パターン照合の結果として各時刻でマッチが検出されるかされないかのみを考えており，本稿のようにマッチの系列としての適切さは評価していない．

一方で，ある時間窓内で Top-k の生起確率を持つマッチを検出する研究も提案されている [7]．しかし，この研究ではパターン照合におけるイベントのスキップを前提としている点が問題となる．例えば，図 1 でパターン $p = \langle \mathtt{a}^+ \ \mathtt{h}^+ \ \mathtt{b}^+ \rangle$ を照合した場合，確率の小さいイベントは無視され $\langle \mathtt{a}_1, \mathtt{h}_4, \mathtt{b}_7 \rangle$ などが Top-1 の生起確率のマッチとして検出される．したがって，本稿で扱うような連続的なイベント生起は検出できない．

3. 準備

パターン照合の入力となる確率的データストリーム PDS 及び問合せパターン p を定める．

3.1 確率的データストリーム

まず，確率的データストリームの要素となる**確率的イベント**を以下に示す．

定義 1 **確率的イベント** e_t は，各イベントシンボル $\alpha \in \Sigma$ に対して生起確率 $P(e_t = \alpha)$ を持つ時刻 t のイベントである．ただし，Σ はイベントシンボルの全集合を示す．また，イベントの生起確率 $P(e_t = \alpha)$ は以下の式を満たす．

$$\forall \alpha \in \Sigma,\ 0 \leq P(e_t = \alpha) \leq 1 \tag{3}$$

$$\forall \alpha, \beta \in \Sigma,\ \alpha \neq \beta \rightarrow P(e_t = \alpha \wedge e_t = \beta) = 0 \tag{4}$$

$$P(\bigvee_{\alpha \in \Sigma} e_t = \alpha) = \sum_{\alpha \in \Sigma} P(e_t = \alpha) = 1 \tag{5}$$

□

簡略化のため，これ以降 $e_t = \alpha$ を α_t で表す．なお，提案手法は文献 [8] のように確率的イベントがマルコフ性を持つ場合にも拡張できるが，議論を簡潔にするために，本稿では時間に関して確率的な独立性を想定する．

確率的イベントを用いて，**確率的データストリーム**を以下のように定義する．

定義 2 **確率的データストリーム** $PDS = \langle e_1, e_2, ..., e_n \rangle$ は，確率的イベントの系列である． □

例えば，図 1 の確率的データストリームは，$\Sigma = \{\mathtt{a}, \mathtt{h}, \mathtt{b}\}$ である確率的イベント e_t の系列 $PDS = \langle e_1, e_2, ..., e_9 \rangle$ として表せる．

3.2 問合せパターン

問合せパターンの記述方法を以下のように定める．

定義 3 入力となるパターン p は以下の文法から生成される．なお，$\alpha \in \Sigma$ であり，ϵ は空イベントを示す．

$$p ::= \alpha \mid \epsilon \mid p\ p \mid p \vee p \mid p^* \mid p^+ \mid (p) \tag{6}$$

□

つまり，本稿ではパターンが一般的な正規表現によって記述されることを想定する．

パターンに適合するマッチとその生起確率を定める．

定義 4 マッチ m は，ユーザが指定したパターン p に適合するイベントシンボルの系列である．マッチの生起確率 $P(m)$ は以下の式で求める．

$$P(m) = P(\bigwedge_{\alpha_i \in m} \alpha_i) = \prod_{\alpha_i \in m} P(\alpha_i) \quad (7)$$

□

4. 情報利得に基づくパターン照合

適切なマッチを検出するために，まずパターン照合のセマンティクスを考える．その後，本稿で扱うパターン照合の問題定義を示す．

4.1 パターン照合のセマンティクス

単純なセマンティクスとして，あるしきい値 τ 以上の情報利得のマッチを検出するというものを考える．式 (2) に示したように，情報利得が 0 以上のときマッチは適切だと判定できる．しかし，現実的には関数 $\theta(m)$ を適切に定めるのは難しいため，0 が適切さを判定するための妥当な値になるとは限らない．そこで，情報利得のしきい値 τ を導入し，以下の式でマッチを出力するか決める．

$$IG(m) = \log_2 \frac{P(m)}{\theta(m)} \geq \tau \quad (8)$$

更に，処理を単純にするために文献 [9] で提案されている *no-overlap* セマンティクスを適用する．no-overlap セマンティクスは，出力されるマッチが同じタイムステップのイベントを持たない，つまりマッチが時間的に重ならないことを要求する．例として，**図 2** のマッチを考える．$\{m_1, m_2, m_3\}, \{m_2, m_3, m_4\}, \{m_3, m_5\}$ のマッチは互いに重なっているため，これらは同時に検出されない．つまり，検出されうるマッチの組合せは以下の 11 通りとなる．

$$\emptyset, \{m_1\}, \{m_2\}, \{m_3\}, \{m_4\}, \{m_5\}, \{m_1, m_4\},$$
$$\{m_1, m_5\}, \{m_2, m_5\}, \{m_4, m_5\}, \{m_1, m_4, m_5\}$$

no-overlap セマンティクスは出力されるマッチの組合せを強く制限するが，実世界におけるイベント生起は時間的に分離しているのが普通であるため，検出に関して問題はないと考える．例えば「部屋 A に留まっている」というイベント生起を考えたとき，このイベントが時間的に重なることはないため，no-overlap セマンティクスでもマッチを適切に検出できる．

4.2 問題定義

no-overlap セマンティクスを適用した上で，情報利得の和が最大となるようマッチを検出することを目指す．

match	time step									$IG(m)$
	1	2	3	4	5	6	7	8	9	
m_1	a									1.32
m_2	a	a	a							3.64
m_3	a	a	a	a	a	a	a	a	a	3.97
m_4		a	a							2.32
m_5							a	a	a	3.32

図 2 マッチと情報利得

no-overlap セマンティクスを用いることで検出されるマッチの組合せを制限できるが，次はいずれの組合せを検出すべきかが問題となる．そこで，本稿では単純に，最も少ない数のマッチで情報利得の和が最大となる組合せを検出する．上述の例について続けて考える．情報利得の和が最大となるのは $\{m_2, m_5\}, \{m_1, m_4, m_5\}$ の二つである．しかし，m_1 と m_4 は m_2 を分割したものであり，あえてこれらを別々に検出する意義はない．したがって，少ない数のマッチで同じ情報利得を持つ $\{m_2, m_5\}$ をパターン照合の結果として出力する．

本稿におけるパターン照合の問題定義を示す．

定義 5 確率的データストリーム PDS, パターン p, 関数 θ, 情報利得のしきい値 τ が入力されたとき，以下の式を満たすマッチ集合 M を検出する．なお，$\mathrm{ts\_overlap}(\mathrm{m}_1, \mathrm{m}_2)$ は m_1 と m_2 の時間的なオーバラップを示す述語である．

$$\begin{aligned} &\text{maximize} \quad \Sigma_{m \in M} IG(m) \\ &\text{subject to} \quad \forall m \in M, IG(m) \geq \tau \\ &\qquad\qquad \forall m_1, m_2 \in M, \neg \mathrm{ts\_overlap}(\mathrm{m}_1, \mathrm{m}_2) \end{aligned} \quad (9)$$

ただし，M が複数考えられる場合はマッチの個数 $|M|$ が最小のものを検出する． □

5. 照合アルゴリズム

最も単純な方法は，全てのマッチを一度検出し情報利得が最大となる組合せを探すというものであるが，この方法は明らかに効率が悪い．クリーネ閉包を含む場合マッチの数は多項式関数的に増えるため，入力されるデータストリームのサイズによってはマッチの検出だけで時間がかかる．加えて，候補となるマッチの組合せは指数関数的に増えるため，全ての組合せを列挙することは難しい．

そこで，定義 5 に基づいてパターン照合を行うとき，集合 M を一つの系列として表せる点に注目する．検出される集合 M は no-overlap セマンティクスを満たすため，任意のマッチ $m \in M$ は同じタイムステップのイベントを持たない．したがって，マッチの間に Σ 中の任意のイベントを表すイベントシンボル “.” を入れることで，集合 M を一つの系列に変換できる．例えば，図 2 で no-overlap を満たすマッチの組合せを考えると，各組合せは**図 3** のように系列として表せる．

sequence	time step									matches
	1	2	3	4	5	6	7	8	9	
π_1	.	.	.	.	.	.	.	.	.	$\emptyset$
π_2	a	.	.	.	.	.	.	.	.	$\{m_1\}$
π_3	a	a	a	.	.	.	a	a	a	$\{m_2, m_5\}$
π_4	a	a	a	.	.	.	a	a	a	$\{m_1, m_4, m_5\}$
π_5	a	a	a	a	a	a	a	a	a	$\{m_3\}$

図 3 マッチ集合と対応する系列

つまり，情報利得が最大となる集合 M を検出するという問題は，情報利得が最大となる系列 $\langle(. \vee p)^*\rangle$ を検出する問題として考えられる．図 3 にあるように，系列 π 中ではマッチ m 若しくは任意のイベント “.” が連続して現れている．これは正規表現 $\langle(. \vee p)^*\rangle$ として表せる．したがって，入力されたパターン p に対して情報利得が最大となる系列 $\langle(. \vee p)^*\rangle$ を検出し，系列に含まれるマッチ集合を出力することで解は求まる．

なお，系列の情報利得は，系列に含まれるマッチの情報利得の和とする．言い換えれば，任意のイベント “.” が持つ情報利得は 0 とする．任意のイベントが生起する確率は常に 1 であるため，情報量 $-\log_2 P(.)$ も常に 0 である．したがって，任意のイベント “.” を検出することで得られる情報は何もないと考えるのが妥当である．

5.1 Viterbi アルゴリズムに基づく照合

Viterbi アルゴリズム [4] は，動的計画法に基づき，状態遷移図における最尤の遷移系列を求めるアルゴリズムである．特に隠れマルコフモデルでよく用いられており，観測系列が得られる確率を最大化する，マルコフモデル上での遷移系列の検出に利用される．本稿では，入力された確率的データストリームとパターンに対して，情報利得を最大化する系列の検出に用いる．

パターン p に対する照合を考えるため，状態遷移図として $\langle(. \vee p)^*\rangle$ に対応する ϵ-NFA（非決定性有限オートマトン）を使用する．ϵ-NFA の生成は以下の 4 ステップで行う．

(1) パターン p に対応する ϵ-NFA をトンプソンの構築法 [10] を用いて生成する．
(2) ϵ-NFA を等価な DFA（決定性有限オートマトン）に変換し，Hopcroft 法 [5] により最小化する．
(3) 全最終状態から初期状態に向け ϵ-遷移を追加し，最終状態を初期状態のみとする．
(4) “.” で初期状態自身に移る遷移を加える．

つまり，パターン p に対応する DFA に対して，照合を繰り返すために初期状態に移る ϵ-遷移と，“.” による遷移とを加えることで生成する．例えば，パターン $p = \langle \mathtt{a}^+\rangle$ が入力されたとき，対応する ϵ-NFA は**図 4** となる．

情報利得が最大となる系列を Viterbi アルゴリズムによって検出する．例として，図 1 の確率的データストリーム PDS，パターン $p = \langle \mathtt{a}^+\rangle$，$\theta = 0.2^{|m|}$，しきい値 $\tau = 2$ が入力された場合を考える．Viterbi アルゴリズムの処理は**図 5** に基づいて行われ，**図 6** に示す過程で系列が生成される．なお，系列は大括弧 “[]” で表し，系列中のマッチは山括弧 “⟨⟩” で区別する．また，下線は各状態で情報利得が最大の系列を示す．開始時点（initial）では初期状態 0 のみが空の系列を持つよう初期化する．タイムステップ 1 では，状態 0 からの遷移が行われるため，状態 0 には $[._1]$ が，状態 1 には $[\langle \mathtt{a}_1\rangle]$ が到達する．更に，状態 1 から 0 への ϵ-遷移があるため，$[\langle \mathtt{a}_1\rangle]$ は最終状態 0 にも到達する．しかし，このとき受理される $\langle \mathtt{a}_1\rangle$ の情報利得は $\tau = 2$ よりも小さいため，この系列は破棄される．タイムステップ 2 でも同様の処理が行われ，各状態に二つの系列が到達する．今興味があるのは情報利得が大きい系列のみであるため，各状態で情報利得が大きい方の系列のみを保持する．タイムステップ 3 以降も同様の処理を行うと，最終的に以下の系列が受理状態 0 に到達する．

$$[\langle \mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3\rangle, .4, .5, .6, \langle \mathtt{a}_7, \mathtt{a}_8, \mathtt{a}_9\rangle]$$

したがって，マッチ $\langle \mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3\rangle, \langle \mathtt{a}_7, \mathtt{a}_8, \mathtt{a}_9\rangle$ を結果として出力する．

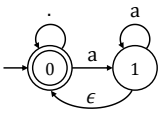

図 4 $\langle(. \vee \mathtt{a}^+)^*\rangle$ に対応する ϵ-NFA

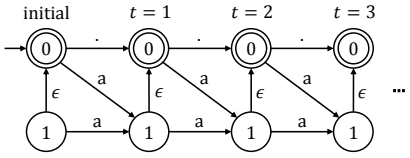

図 5 Viterbi アルゴリズムの処理過程

t	state	
	0	1
initial	{[]}	{}
1	$\{[\underline{.1}]\}$	$\{[\underline{\langle \mathtt{a}_1\rangle}]\}$
2	$\{[.1, .2], [\underline{\langle \mathtt{a}_1, \mathtt{a}_2\rangle}]\}$	$\{[.1, \langle \mathtt{a}_2\rangle], [\underline{\langle \mathtt{a}_1, \mathtt{a}_2\rangle}]\}$

図 6 系列の生成過程

6. 評価実験

実データを用いて検出性能を，人工データを用いて処理速度を評価する．実データには Lahar プロジェクト [8] で公開されている屋内位置の確率的データストリームを使用する．実データは**図 7** のように屋内構造をグラフで表しており，被験者が各ノードにいる確率と実際にその被験者がいた正解ノードの情報を一秒毎に持つ．つまり，パターン p を与えたとき，正解ノードの情報を用いることで正解と

Vol.2016-DBS-163 No.5
Vol.2016-IFAT-123 No.5
2016/9/13

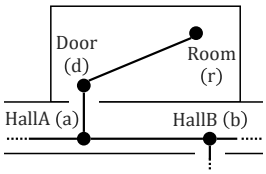

図 7 実データが持つグラフ構造の一部

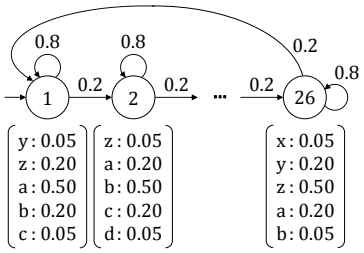

図 8 人工データ生成のためのマルコフモデル

なるマッチも検出できる．人工データには**図 8** に示すマルコフモデルを用いてイベント数 100 万の確率的データストリームを生成した．生成は初期状態 1 から始まり，各遷移確率に応じて状態を遷移する．各状態では五つのイベントシンボルを持つ確率的イベントを出力する．なお，イベントシンボルとしてアルファベット $\Sigma = \{\mathtt{a}, \mathtt{b}, ..., \mathtt{z}\}$ を使用したため，シンボルの総数は 26 個である．

6.1 検出性能の評価

適合率・再現率・F 値を用いて検出性能を評価する．ただし，確率的なパターン照合では検出結果と正解データが完全一致することは稀であるため，二つのマッチがどれだけオーバラップするかを示す重複度を基に各指標を定める．

$$\mathrm{overlap}(m, c) = \frac{\min(m.t_e, c.t_e) - \max(m.t_s, c.t_s) + 1}{\max(m.t_e, c.t_e) - \min(m.t_s, c.t_s) + 1} \quad (10)$$

適合率は検出の正確性，つまりマッチが正解データとどれだけ一致するかを表す．本稿では，適合率を正解データに対するマッチの重複度の平均として定義する．

$$\mathrm{precision}(M, C) = \frac{1}{|M|} \sum_{m \in M} \max_{c \in C}(\mathrm{overlap}(m, c)) \quad (11)$$

再現率は検出された正解データの割合であり，検出の網羅性を表す．適合率と同様に，再現率はマッチに対する正解データの重複度の平均として定める．

$$\mathrm{recall}(M, C) = \frac{1}{|C|} \sum_{c \in C} \max_{m \in M}(\mathrm{overlap}(m, c)) \quad (12)$$

F 値は適合率と再現率の調和平均であり，検出の全体的な性能を表す．本稿では適合率と再現率を一対一の割合で評価し，F 値を以下の式で計算する．

$$\mathrm{F}(M, C) = \frac{2 \cdot \mathrm{precision}(M, C) \cdot \mathrm{recall}(M, C)}{\mathrm{precision}(M, C) + \mathrm{recall}(M, C)} \quad (13)$$

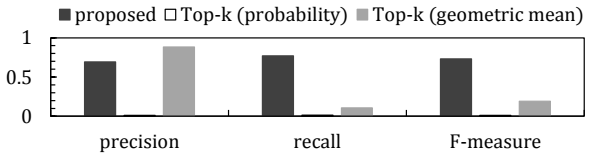

図 9 $p = \langle \mathtt{Door}^+\ \mathtt{Room}^+\ \mathtt{Door}^+ \rangle$ に対する適合率・再現率・F 値

比較手法として，マッチの生起確率に基づく Top-k 問合せと，各イベントの生起確率の幾何平均に基づく Top-k 問合せの二つを用いる．なお，k の値には提案手法により検出されたマッチの数を用いる．生起確率・幾何平均共に，Top-k 問合せの解は Viterbi アルゴリズムを拡張することで求められる．

提案手法を用いることで，適合率・再現率・F 値をバランスよく高められる．**図 9** に $p = \langle \mathtt{Door}^+\ \mathtt{Room}^+\ \mathtt{Door}^+ \rangle, \theta(m) = 0.1^{|m|}, \tau = 40$ を入力した際の実験結果を示す．この入力のとき，正解データ 9 個に対して 10 個のマッチが検出された．各正解データは 20 から 40 秒ほどにわたって生起しているが，生起確率による Top-k 問合せでは 3 から 4 秒ほどのマッチしか検出できないため，適合率・再現率共に非常に小さい．幾何平均を用いると適合率は大きくなるが，全 9 個の正解データのうち一つに対応するものしか検出できていないため，再現率は小さいままである．一方，情報利得に基づくマッチの評価と，no-overlap セマンティクスによるマッチの選別により，提案手法は適合率・再現率共に大きい．特に再現率は比較手法に比べ大きく，全正解データに対応するマッチを検出できたのは提案手法のみである．

最後に，この実験における $\theta(m)$ 及び τ の設定方法について述べる．$\theta(m)$ は，生データが持つ確率の値から設定した．Lahar データセットにおいて各イベントシンボルの確率は，0.2 より大きいものと 0.01 未満の小さいものとにおおよそ二分されている．そこで，マッチの確率の推定として，各イベントシンボルの確率が 0.1 である場合を想定した $\theta(m) = 0.1^{|m|}$ を使用した．なお，しきい値 τ により出力するマッチを調整できるため，$\theta(m)$ はマッチの検出漏れがないよう小さめの推定値が算出されるようにした．一方で，τ の値は実験的に決定した．今回は，τ の値を 0 から 100 まで変化させ，F 値が最大となった $\tau = 40$ を実験で使用した．実際の τ による適合率・再現率・F 値の変化を**図 10** に示す．τ を高くすることで余計なマッチが破棄され適合率は大きくなるが，高すぎる τ は重要なマッチまで破棄してしまうため再現率が小さくなる．したがって，適合率と再現率のバランスが良い 25 から 50 あたりで F 値は最大となっている．

6.2 処理速度の評価

提案手法は，入力となるパターン p が複雑になっても効率よく照合できる．**図 11** にパターンの長さ $|p|$ を 1 から

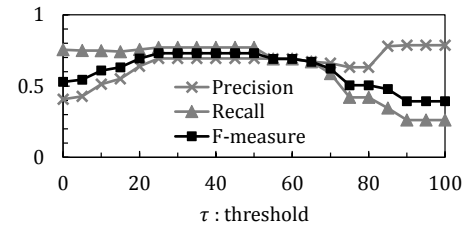

図 10 τ の変化に対する適合率・再現率・F 値

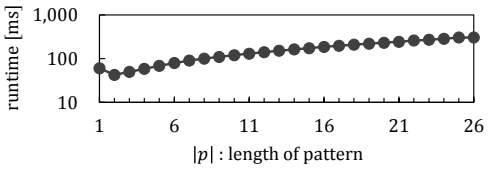

図 11 パターンの長さ $|p|$ を変化させた際の実行時間

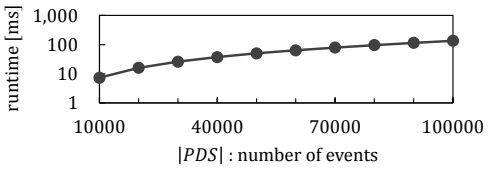

図 12 イベント数 $|PDS|$ を変化させた際の実行時間

26 まで変化させた際の実行時間を示す．パターンの長さは，$\langle \mathtt{a}^+ \rangle, \langle \mathtt{a}^+\ \mathtt{b}^+ \rangle$ のように，イベントを順に追加することで調整した．なお，確率的データストリームのイベント数は 50,000 であり，$\theta(m) = 0.1^{|m|}$ 及び $\tau = 0$ を用いた．パターンの長さは ϵ-NFA の遷移数に影響を与えるが，Viterbi アルゴリズムの実行時間において遷移数は線形にしか影響しないため，パターンが長さが実行時間に与える影響は小さくなっている．

また，提案手法は，入力となる確率的データストリームの持つイベント数が多くても効率よく照合できる．**図 12** にイベント数 $|PDS|$ を 10,000 から 100,000 まで変化させた際の実行時間を示す．なお，他の入力には $p = \langle \mathtt{a}^+\ \mathtt{b}^+\ \mathtt{c}^+ \rangle$，$\theta(m) = 0.1^{|m|}$ 及び $\tau = 0$ を用いた．Viterbi アルゴリズムにおいて，データストリームのイベント数は系列を更新する回数に等しい．つまり，イベント数は実行時間に線形の影響しか与えない．したがって，イベント数が増えても処理の効率は変わらず，毎秒数十万イベントのスループットを達成している．

7. おわりに

確率的データストリーム上でのパターン照合において，適切なマッチを検出するための評価指標として情報利得を提案した．マッチの情報利得を定め，no-overlap セマンティクスに基づきパターン照合の問題定義を定めた．また，no-overlap セマンティクスを用いたとき，Viterbi アルゴリズムにより効率的に解を検出できることを示した．実データを用いた実験により提案した情報利得が既存の評価指標（生起確率・幾何平均）よりも優れていることを示し，人工データを用いた実験によって実行時間においても優れていることを示した．

今後の課題としては，提案手法のリアルタイム処理への拡張，複数のデータストリームにまたがったパターンへの適用が挙げられる．また，今回はマッチの枝刈りにしきい値 τ を用いたが，Top-k 問合せによるマッチの枝刈りも考えられる．他にも，文献 [9] において no-overlap セマンティクスよりも優れたものとして提案されている use-and-throw セマンティクスを提案手法に適用することも今後の課題である．

謝辞 本研究の一部は科研費（16H01722, 26540043）および JST COI プログラムによる．

参考文献

[1] Aggarwal, C. C. and Yu, P. S.: A Framework for Clustering Uncertain Data Streams, *2008 IEEE 24th ICDE*, pp. 150–159 (2008).
[2] Chen, L., Nugent, C. and Wang, H.: A Knowledge-Driven Approach to Activity Recognition in Smart Homes, *IEEE TKDE*, Vol. 24, No. 6, pp. 961–974 (2012).
[3] Cormode, G. and Garofalakis, M.: Sketching Probabilistic Data Streams, *Proc. 2007 ACM SIGMOD*, pp. 281–292 (2007).
[4] Forney, Jr., G. D.: The Viterbi Algorithm, *Proc. IEEE*, Vol. 61, No. 3, pp. 268–278 (1973).
[5] Hopcroft, J. E.: An N Log N Algorithm for Minimizing States in a Finite Automaton, Technical report, Stanford University (1971).
[6] Jin, C., Yi, K., Chen, L., Yu, J. X. and Lin, X.: Sliding-window Top-k Queries on Uncertain Streams, *Proc. VLDB Endow.*, Vol. 1, No. 1, pp. 301–312 (2008).
[7] Li, Z., Ge, T. and Chen, C. X.: ε-Matching: Event Processing over Noisy Sequences in Real Time, *Proc. 2013 ACM SIGMOD*, pp. 601–612 (2013).
[8] Ré, C., Letchner, J., Balazinska, M. and Suciu, D.: Event Queries on Correlated Probabilistic Streams, *Proc. 2008 ACM SIGMOD*, pp. 715–728 (2008).
[9] Santini, S.: Querying Streams using Regular Expressions: Some Semantics, Decidability, and Efficiency Issues, *VLDB J.*, Vol. 24, No. 6, pp. 801–821 (2015).
[10] Thompson, K.: Programming Techniques: Regular Expression Search Algorithm, *Commun. ACM*, Vol. 11, No. 6, pp. 419–422 (1968).
[11] Tran, T. T. L., Peng, L., Diao, Y., McGregor, A. and Liu, A.: CLARO: Modeling and Processing Uncertain Data Streams, *VLDB J.*, Vol. 21, No. 5, pp. 651–676 (2012).
[12] Yin, J., Yang, Q. and Pan, J. J.: Sensor-Based Abnormal Human-Activity Detection, *IEEE TKDE*, Vol. 20, No. 8, pp. 1082–1090 (2008).
[13] Zhang, Q., Li, F. and Yi, K.: Finding Frequent Items in Probabilistic Data, *Proc. 2008 ACM SIGMOD*, pp. 819–832 (2008).